Disambiguation of Emotion Annotations by Contextualizing Events in Plausible Narratives

Johannes Schäfer, Roman Klinger
Fundamentals of Natural Language Processing, University of Bamberg, Germany
{johannes.schaefer,roman.klinger}@uni-bamberg.de

Abstract

Ambiguity in emotion analysis stems both from potentially missing information and the subjectivity of interpreting a text. The latter did receive substantial attention, but can we fill missing information to resolve ambiguity? We address this question by developing a method to automatically generate reasonable contexts for an otherwise ambiguous classification instance. These generated contexts may act as illustrations of potential interpretations by different readers, as they can fill missing information with their individual world knowledge. This task to generate plausible narratives is a challenging one: We combine techniques from short story generation to achieve coherent narratives. The resulting English dataset of Emotional BackStories, EBS, allows for the first comprehensive and systematic examination of contextualized emotion analysis. We conduct automatic and human annotation and find that the generated contextual narratives do indeed clarify the interpretation of specific emotions. Particularly relief and sadness benefit from our approach, while joy does not require the additional context we provide.

Keywords: Emotion Analysis, Context, Synthetic Data, Story Generation, Human Annotation

1. Introduction

When interpreting event descriptions, ambiguity in emotion analysis can arise for various reasons. Studies have attempted to explain low inter-annotator agreement by differences in demographics among annotators, which can contribute to varied interpretations in subjective tasks (Wan et al., 2023; Mieleszczenko-Kowszewicz et al., 2023; Plaza-del Arco et al., 2024; Sun et al., 2025, i.a.). Factors such as age, cultural background, and personal experiences may influence the interpretation of emotion responses to an event.

However, what has received no attention so far is the impact of variations in the textual context of event descriptions, which can also significantly shape emotion interpretations and responses. For instance, given the following text from the dataset[1] we present in this paper: *"The loudspeaker suddenly malfunctioned and went silent."*, a prevalent interpretation may be that this event causes surprise. However, the context of the event may change the distribution of possibly experienced emotions as shown in Table 1. In the case of guilt, the added narrative establishes a profound sense of personal responsibility and the anticipated negative consequences of actions, which are then triggered by the event. Relief emerges from including a backstory that portrays a situation initially fraught with anxiety and possibly negative consequence beyond one's control, suggesting that these are re-

[1]We provide our annotated dataset and code on https://www.uni-bamberg.de/en/nlproc/resources/emotional-backstories/.

Guilt: A close friend entrusted me with setting up the sound system for a charity event. I faced a daunting task, as the previous technician had left the equipment in disarray. With time running out, I took a shortcut to meet the deadline, skipping some crucial safety checks. Just before the event started, my friend reminded me of its significance and the many people counting on its success. *The loudspeaker suddenly malfunctioned and went silent.*
Relief: I was tasked with giving a presentation to a large crowd. The sound system malfunctioned, amplifying my voice to an ear-piercing level. The sound technician ignored the problem and chatted with someone. The audience covered their ears and looked at me with discomfort. *The loudspeaker suddenly malfunctioned and went silent.*
Fear: I arrived at a remote wilderness survival training camp, where the instructors emphasized the importance of following loudspeaker instructions for safety. The instructors warned us about the toxic waste site nearby and explained that the loudspeaker would alert us to any changes in air quality. During the first exercise, I struggled to navigate the challenging terrain, but the loudspeaker provided crucial guidance, helping me stay on track. I completed a difficult obstacle course, relying heavily on the loudspeaker's instructions to avoid hazards and find the safest route. *The loudspeaker suddenly malfunctioned and went silent.*
Pride: I spent the entire morning upgrading the sound system with a new backup system to prevent technical issues. The event organizer informed me that the conference was running 30 minutes behind schedule, giving me extra time to test the new backup system. I used the extra time to run a series of tests on the sound system, trying to simulate potential failures. The keynote speaker began to talk, and the sound system was working flawlessly, but I was still waiting for a real test of the new backup system. *The loudspeaker suddenly malfunctioned and went silent.*

Table 1: Example narratives for different emotions.

solved by the event. Fear is framed by a narrative that expresses the unpredictability of a challenging circumstance, emphasizing a sense of uncertainty and the immense effort required to navigate the unknown, reinforcing an unpleasant loss of control based on the event. Lastly, pride is evoked by incorporating a backstory that highlights proactive efforts and successful task management, illustrating how these accomplishments will be recognized by others as a result of the event. These examples demonstrate that backstories elicit cognitive appraisals for each response, which lead to emotion interpretation differences in the event. We assume that readers of ambiguous events similarly fill this context based on their world knowledge.

While there have been studies on emotion classification under the paradigm of perspectivism (Milkowski et al., 2021; Suzuki et al., 2022; Kazienko et al., 2023; Troiano et al., 2023, i.a.), we are not aware of any work that studies the importance of context to disambiguate emotion analysis in a controlled manner.[2] Ambiguity in the interpretation of event descriptions viewed in isolation can stem from missing essential information, which can also lead to uncertainty in emotion responses. Furthermore, not all event descriptions elicit distinct emotions; some are devoid of specific emotion, and thus can be filled only through narratives. We approach disambiguation by generating different contexts as sequences of preceding events, thereby examining how various backstories can have different effects on the emotion interpretation of that event. Our proposed systems generate event chains by prompting a large language model (LLM) in different settings, with the goal to understand whether these contexts lead to a higher agreement in human annotation of emotions.

In our experiments, we first assess the quality of our generated data, to ensure that it is suitable for an evaluation of human annotation and automatic predictions, which is our main goal. Hence, our paper is guided by the following research questions:

1. Does context generation through iterative prompting with story planning enhance narrative coherence or is a one-step prompting approach sufficient?
2. Do generated contexts enhance clarity in human annotation of emotions in events?
3. Can systems recognize the contextual disambiguation in emotion analysis of events?

To answer the questions, we construct a dataset (Emotional BackStories, EBS) as the foundation to study the influence of contextual information on emotion prediction. Our work is situated at the intersection of explainable artificial intelligence (XAI) and perspectivism. Our contribution fits to XAI, because the generated context functions as an explanation of a possible textual interpretation. It is related to perspectivism, because multiple such contexts are reasonable.

In the remainder of this paper, we review related work in Section 2, present our data generation and analysis methods in Section 3, and analyze our experimental results in Section 4.

[2] Emotion recognition in conversations does consider context, but not to disambiguate otherwise unclear utterances in a controlled manner (Hu et al., 2021, i.a.).

2. Related Work

The novel task of generating backstories that evoke specific emotions in events intersects with multiple research fields, which we review in this section.

2.1. Contextual Influences on Event Interpretations

The interpretation of text is significantly shaped by its context. Das et al. (2011) explore the dynamics of emotions in their analysis of event chains, with the goal of finding the correct temporal sequencing of events. They focus on identifying the sentiments and emotions associated with events and utilize these as features to uncover contextual relationships. Mostafazadeh et al. (2016) introduce the ROCstory dataset, which comprises crowd-sourced narratives, each consisting of a series of five events. This corpus is intended to assist in story completion tasks, where a system must determine the most fitting event to follow the first four provided events. Park et al. (2020) conduct sentence-level emotion analysis on this dataset, although they do not take prior contextual influences into account. The GLUCOSE dataset introduced by Mostafazadeh et al. (2020) further enhances understanding of complex dependencies between events through commonsense inferences; however, it does not specifically address the emotions being evoked in these contexts.

2.2. Emotion Categorization in Context

Emotion analysis seeks to identify emotion states elicited in readers or authors of texts. The foundational frameworks by Ekman (1972) and Plutchik (2001) provide categorization methodologies. Mohammad (2012) created emotion-labeled social media corpora, while Troiano et al. (2023) emphasized implicit emotion cues linked to appraisal theories. Their analysis of event descriptions without context highlights a gap our study addresses by focusing on contextual influences. Notably, Etienne et al. (2022) emphasize the intricacies of emotion expression across diverse texts, and Étienne et al. (2024) predict various modes of emotion expression, emphasizing the complexity that this

context brings to emotion annotation. Chatterjee et al. (2019) and Wemmer et al. (2024) advocate for considering prior context in dialog systems to refine emotion interpretations. The dynamic interplay between event components and emotions is also acknowledged by Cortal et al. (2023). Additionally, Labat et al. (2024) investigate emotion trajectories in customer service dialogues, emphasizing that emotions may shift with each utterance. Our work enhances these studies with an in-depth analysis of contextual influences on emotion analysis in a controlled dataset tailored for this purpose.

2.3. Story Generation Techniques

Story generation methodologies central to our study involve techniques designed to enhance narrative quality and coherence. Razumovskaia et al. (2024) emphasize the benefits of story planning, while Ma et al. (2023) examine how integrating structured knowledge can improve coherence. Xie and Riedl (2024) and Yu et al. (2023) demonstrate effective planning and diversification strategies, and Hosseini et al. (2024) focus on generating high-variance datasets for natural language inference tasks. Additionally, Chung et al. (2023) analyze the balance between accuracy and diversity of generated data. Yang and Jin (2024) address challenges in establishing effective criteria for evaluating story quality, while Long et al. (2024) discuss the need for human intervention to maintain data faithfulness and diversity. Furthermore, Eigenschink et al. (2023) and Li et al. (2023) emphasize the importance of generating synthetic data that adheres to realism, diversity, and coherence. These varied insights guide the generation approach in our setting with regard to narrative quality and relevance.

2.4. Backstory Generation

Clark and Sood (2022) explore character backstories in game design with a multi-step approach to develop engaging narrative complexities. Moon et al. (2024) produce realistic persona backstories using open-ended prompts, while Stricker and Paroubek (2024) suggest including specific instructions in prompts in order to create user backstories for enhanced conversational realism. Furthermore, Jung et al. (2023) focus on generating preceding dialogue turns to augment user-system interactions with a focus on cohesive alignment.

Overall, the literature emphasizes the necessity of coherent, contextually diverse narratives. Existing research tends to generate free-form stories or focus on story completion tasks, where the emphasis is on crafting endings for pre-existing narratives. In contrast, our study aims to generate backstories that influence specific events, necessitating a novel reverse construction method. Insights into prompt setups, stepwise approaches, content diversification, story planning, and evaluation criteria for narrative quality further inform our methodology.

3. Generation of Event Chains

In order to systematically perform contextualized emotion analysis, we generate short stories that contain narratives according to specific requirements with an LLM, considering various settings. We now define the underlying concepts, outline the procedural steps of our generation approaches, and present our evaluation methods.

3.1. Definitions

Emotion Modeling. In our study, emotion analysis of text refers to modeling a person's interpretation of the events they describe. Following Troiano et al. (2023), we use a fine-grained emotion model consisting of $n=13$ categories: $\mathbb{E}=\{e_1,\ldots,e_n\}=\{$anger, boredom, disgust, fear, guilt, joy, pride, relief, sadness, shame, surprise, trust, no-emotion$\}$. The rationale for selecting these categories stems from their grounding in appraisal processes, which provide vital insights into emotion responses. We suspect that relevant appraisal processes may be articulated within the narrative contexts we aim to generate. Moreover, by employing a larger set of emotions, we allow for a fine-grained analysis that can lead to richer and more diverse texts. Broader analyses can be performed by grouping of some finer categories into overarching classifications, such as distinguishing between positive and negative emotions.

Event Chains. We generate short stories as event chains consisting of five distinct sentences $c=\{s_1,...,s_5\}$, which is in line with data in the corpus ROCstory (Mostafazadeh et al., 2016). Each sentence contains a description of a singular event from the perspective of a person experiencing it. The five events in a chain are organized chronologically to form a coherent narrative. We consider the first four events in a chain as the backstory $b=\{s_1,\ldots,s_4\}$ and analyze its influence on the emotion being evoked in the event experiencer in the concluding event $e(b\cdot s_5)$. To facilitate a comprehensive analysis of diverse emotion responses, we associate multiple distinct backstories with the same concluding event s_5. For our dataset, we aim to include precisely one backstory for each emotion category for each concluding event: $\mathbb{B}(s_5)=\{b_i \mid e(b_i\cdot s_5)\in\mathbb{E}\}$ with $i=1,\ldots,n$.

3.2. Generation Approaches

We generate diverse event descriptions (the set of events s_5) and then compare three methods

Step	Input	Summarized Prompt	Output
I.	Event type: *Competition*, Event type object: *communication tools*.	Prompt 1: Describe event given type and object.	*The loudspeaker suddenly malfunctioned and went silent.*
II.	Event: *The loudspeaker suddenly malfunctioned and went silent.*, Emotion: *Guilt*.	Prompt 2: Construct backstory for emotion and event.	I had been tasked with testing the loudspeaker system before the big event. My supervisor warned me that a malfunction would be disastrous for the company's reputation. I skipped the recommended final check to grab a quick lunch before the event started. The event host began the ceremony, and the loudspeaker was working perfectly, filling me with temporary relief. *The loudspeaker suddenly malfunctioned and went silent.*

Table 2: Steps for target event generation and backstory generation (Baseline method) on an example.

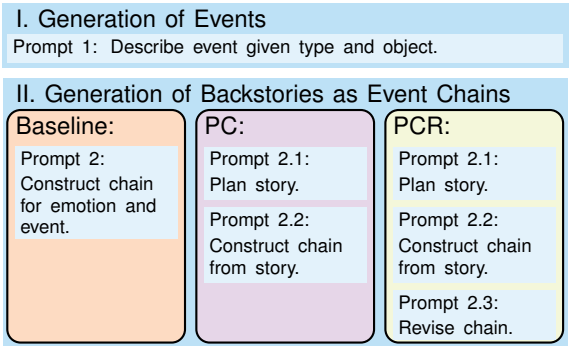

Figure 1: Overview of our LLM-based data generation framework for event chains according to three different methods. The prompts used in each case are shown summarized – for full text prompts see Appendix B.

for generating contexts for each of them, as visualized in Figure 1. The generation process, based on the instruction-tuned Llama-3.1-70B-Instruct model (Meta, 2024), is guided via prompting (see Appendix B for the used prompts).

I. Generation of Events. The initial step generates a set of event descriptions, which act as instances to be interpreted in a real-world scenario. We ensure diversity by adopting the attribute-based strategy proposed by Yu et al. (2023), leading to a balanced distribution of topics, as mentioned in the prompt (Step I in Table 2; topics: Appendix A). We perform few-shot prompting with ten examples.

II. Generation of Backstories. Given an event description s_5, the second step creates the corresponding backstories $\mathbb{B}(s_5)$. We specify the emotion the resulting event chain should evoke in the prompt, i.e., for every s_5 from the generated set of events, we include $e(b_i \cdot s_5)$ for every $b_i \in \mathbb{B}(s_5)$ with $i{=}1, \ldots, n$. Guiding principles are this specific influence on the evoked emotion and narrative coherence. We generate a different backstory for each emotion category based on a given event. To address this challenge within a forward-generation LLM, we convert the backward generation task into a forward generation approach by embedding specific instructions into the prompt.

We compare three approaches (Figure 1): In the Baseline, all requirements are specified within a single prompt (Step II in Table 2). In our Plan–Construct (PC) chain-of-thought method, we perform story generation decomposition (inspired by Clark and Sood, 2022) and story planning (following Razumovskaia et al., 2024). The method Plan–Construct–Revise (PCR) subsequently revises the chain in consideration of the two previous outputs.

3.3. Evaluation Methods

We aim at understanding if including the generated contexts allows humans and automatic classification to estimate the associated emotion more consistently. At the same time, the created narratives shall be coherent.

Coherence. We evaluate coherence with a zero-shot shuffle test (Laban et al., 2021) using an LLM to compute the sequence likelihood of event chains (Appendix C.2). The resulting scores are compared against various permutations in which the order of events is shuffled. A high coherence score (total range from 0 to 1) is assigned to the original event chain if it ranks favorably among its shuffled permutations.

Human Annotation. We analyze generated events as well as the impact of backstories on evoked emotions by conducting human annotation studies. We ask crowd workers to annotate emotions as well as to assess the quality of a sample (3 annotators per instance, see Appendix E for the setup).[3] Given an event chain (or just one event description), annotators have to select the

[3]We use Prolific (https://www.prolific.com/).

most prominent emotion the author of the text presumably felt in the end. Events are additionally annotated on a Likert scale from 1 to 5 regarding vagueness and plausibility, as well as whether they are assumed to be written by a human or by an AI. An important aspect in the evaluation is that we do not only evaluate the average agreement, but particularly aim to identify the cases where a reasonable context allows for an improved annotation.

For the human annotation, we sample a set of 10 events from our dataset (one from each event type category) and collect 39 annotations for this sample.[4] Additionally, we sample 90 more events (nine from each event type category) and collect 3 annotations each, so we can evaluate the quality of the generated texts on a sample of 100 events (10% of our dataset). To annotate entire narratives, we collect 780 annotations of event chains, which corresponds to three annotations per 13 backstories, generated by each of the two methods, for each of the 10 sampled events.

Emotion Analysis. To assess the impact of additional context on automatic classification, we zero-shot classify emotions.[5] For each instance (events, backstories, or event chains), we prompt the model with a set of input templates where each includes a specific emotion label (as shown in Table 3) and record emotion category probabilities based on the model's sequence likelihoods. Therefore, we compute $p(e|\{s_1, ..., s_m\})$ for $m=1, \ldots, 5$ and every $e \in \mathbb{E}$. The emotion classification performance is on par with the scores reported by (Troiano et al., 2023) on a comparable corpus with the same emotion categories (our approach achieves .54 F1, vs. .59). Their work employs a fine-tuned RoBERTa-large model (Zhuang et al., 2021; Liu et al., 2019).

4. Analysis

Our dataset is based on 1000 generated event descriptions consisting of 15.4 tokens on average (Appendix C). The instances exhibit a low level of vagueness and are plausible (based on human annotation of 100 sampled event descriptions, vagueness: mean 2.50, $\sigma=1.20$, plausibility: mean 4.32, $\sigma=0.89$). The relatively high standard deviation for vagueness aligns well with our goal of creating a diverse dataset. The annotators tend to attribute authorship to humans (mean 3.50, $\sigma=0.99$) over AI (mean 2.67, $\sigma=1.05$).

We generate 13 backstories for each event, corresponding to the set of emotion categories studied. Consequently, our dataset comprises 13,000 event chains for each of the three methods. The Baseline generates backstories averaging 69.8 tokens, while PC backstories average 49.2 tokens and PCR backstories average 79.2 tokens (Appendix C). Examples are shown in Table 1. We conduct an analysis for lexical diversity, coherence and emotions to answer our research questions in the following subsections.

MT	Prompt Message Template
System	You are an expert in emotion analysis on event descriptions.
User	A person describes their experience as follows: {text_instance} What emotion was evoked in the person at the end? As your response, provide only one label from the emotion set: anger, disgust, fear, guilt, joy, sadness, shame, pride, boredom, surprise, trust, relief, no-emotion.
Assistant	{emotion}

Table 3: Template of prompt messages for our zero-shot emotion analysis on events and event chains. 'MT' refers to the prompt message type as specified in the input for the instruction-tuned LLM.

4.1. Does context generation through iterative prompting with story planning enhance narrative coherence or is a one-step prompting approach sufficient?

A requirement for the analysis of evoked emotions is that the data is of high quality. We first evaluate the coherence scores of the event chains generated by our three methods. The results displayed in Table 4 show that PCR (.84) and PC (.77) both outperform the baseline method (.75): Story planning does contribute to more coherent narratives. In a qualitative analysis we further find that PC and PCR generate more diverse backstories than the baseline (see Appendix D). The variation in story coherence across different emotion categories is relatively low. All scores are on an acceptable level and the content diversity of the generated texts is high (see Appendix C.3). Fear shows the lowest coherence scores for all methods.

Potential Emotion Leakage in Backstories. The generated backstories are designed to avoid explicit references to the targeted emotions. To assess adherence to this aim, we analyze potential emotion leakage. We use ChatGPT (OpenAI, 2025) to compile lists of 25 synonyms for each of our emotion categories and manually refine these to only contain unique terms. We count cases

[4] We collect 39 annotations for each event since we intend to compare these to the annotations of 13 corresponding event chains (with 3 annotations each).

[5] We use Llama-3.1-70B-Instruct (Meta, 2024).

		Mean Coherence of Chains		
		Baseline	PC	PCR
Emotion-Specific Subsets	Anger	.75	.78 +.03	.84 +.09
	Boredom	.73	.76 +.03	.84 +.11
	Disgust	.72	.77 +.05	.84 +.12
	Fear	.67	.72 +.05	.80 +.13
	Guilt	.77	.79 +.02	.86 +.09
	Joy	.76	.78 +.02	.85 +.09
	Pride	.81	.78 −.03	.85 +.04
	Relief	.81	.78 −.03	.85 +.04
	Sadness	.76	.79 +.03	.84 +.08
	Shame	.78	.79 +.01	.86 +.06
	Surprise	.76	.79 +.03	.83 +.07
	Trust	.78	.80 +.02	.85 +.07
	No-Emotion	.69	.74 +.05	.82 +.13
Overall Dataset		.75 $\sigma = .26$	.77 +.02 $\sigma = .25$	.84 +.09 $\sigma = .21$

Table 4: Mean coherence scores for event chains in the datasets generated with different methods. Delta values (marked with +/−) show the difference in comparison to the Baseline.

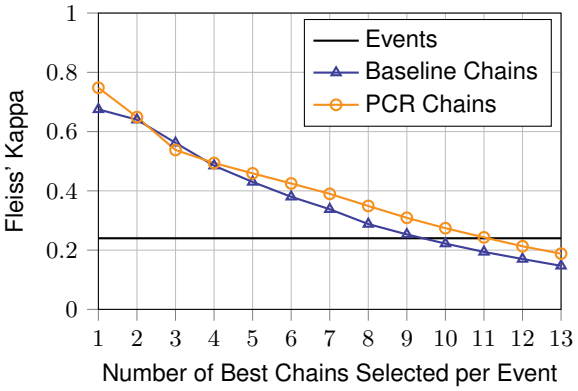

Figure 2: Emotion annotation agreement on events in comparison to Baseline and PCR chains.

where any of these terms or the emotion word itself appear in the backstories as emotion leakage.

We find only a low issue with leakage rates in backstories: 2% of the background stories contain an emotion word in the Baseline (310 out of 13,000), 1% for PC (165 out of 13,000), and 5% for PCR (682 out of 13,000). Paraphrasing these terms or filtering out the instances could mitigate this issue. However, we chose to retain the few instances of emotion leakage in our analysis to maintain a straightforward evaluation process, recognizing that filtering might inadvertently also introduce errors. This approach highlights opportunities for refinement in future research.

$e(b \cdot s_5)$ \ e_A	Anger	Boredom	Disgust	Fear	Guilt	Joy	Pride	Relief	Sadness	Shame	Surprise	Trust	No-Emot.
Anger	5	0	0	5	0	0	0	3	11	1	3	0	2
Boredom	5	0	2	1	0	2	3	0	8	0	1	0	8
Disgust	3	1	6	1	0	4	0	1	2	2	3	1	6
Fear	4	0	1	12	2	0	0	1	1	2	4	0	3
Guilt	3	0	0	3	6	0	2	4	4	3	1	0	4
Joy	1	0	1	0	0	2	7	7	5	1	1	1	4
Pride	3	0	0	0	3	0	10	7	0	0	0	1	6
Relief	3	0	0	5	0	0	1	6	4	0	3	2	6
Sadness	2	0	0	3	0	1	1	1	15	0	3	1	3
Shame	0	1	0	3	0	3	2	0	1	11	3	0	6
Surprise	4	0	1	0	1	3	2	0	7	1	6	2	3
Trust	3	1	0	0	0	1	7	9	2	0	0	2	5
No-Emot.	6	0	0	2	0	2	2	8	2	1	2	0	5
Σ	42	3	11	35	12	18	37	47	62	22	30	10	61

Table 5: Prompted emotion $e(b \cdot s_5)$ vs. annotated emotion e_a on 130 PCR event chains (3 annotations each), 10 per $e \in e(b \cdot s_5)$.

4.2. Do generated contexts enhance clarity in human annotation of emotions in events?

We now assess the impact of our generated contexts on the clarity of emotion interpretations based on results of the human annotation on a sample of our dataset. This sample consists of 10 distinct events, each enriched by one Baseline backstory as well as one PCR backstory for each of the 13 emotion categories studied, i.e. two sets of 130 event chains.

Our goal is to determine whether the introduction of the contexts clarifies the emotion categorization given the event chains for humans compared to when assessing the individual events. Therefore, we calculate the inter-annotator agreement (Fleiss' kappa), which is .24 on events, .15 for Baseline chains and .19 for PCR chains. However, we should take into account that this evaluation encompasses the backstories generated across all the different emotion categories for each event. Since an event may not evoke all possible emotions even when viewed in context, it is reasonable to expect that disambiguation would only be feasible for certain relevant emotion categories. Therefore, we evaluate if there are valuable contexts available to the annotators by analyzing only a selection of the generated chains relevant to each event.

We find that (see Figure 2) for PCR chains, a vast majority – 10 out of the 13 backstories generated for each event – successfully provide informative context that enhances clarity in emotion analysis for annotators. Our approach effectively disambiguates emotion interpretations of given events.

To further understand how humans interpret

emotions within the contextualized events provided in our dataset, we investigate the specific emotion categories assigned during the annotation process. In Table 5, the counts of annotated labels (columns) of the PCR chains generated for the different prompted emotions (rows) are organized. Annotators recognize the prompted emotion most often for fear, pride, sadness, and shame. Notably, certain patterns arise: Chains designed to trigger anger are often annotated as sadness, while those prompted for trust are frequently classified as relief. Additional analysis confirms that such typical overlaps among the annotated emotion categories are evident in our data (see Appendix E). Overall, for most emotions the prompted emotion is the category which is most often annotated by the human annotators. This further reinforces the efficacy of our context generation framework in clarifying emotion interpretations for humans.

4.3. Can systems recognize the contextual disambiguation in emotion analysis of events?

Annotation Correlation. To understand if our automatic emotion analysis aligns with human annotations, we compute the Spearman correlation between the probabilistic prediction score and the proportion of annotators who identified the respective emotion. The overall results reveal a correlation of .47 for events and .37 for chains – both significant ($p < .001$). Further analysis of the correlation shows that events exhibit greater fluctuations across emotion categories, whereas assessments of chains are more consistent (see Appendix F).

The Overall Influence of Backstories in Emotion Analysis. The probability of each emotion category being evoked on average across the 1000 events (see Table 6, under the column labeled E) reveals a typical distribution of emotions associated with general descriptions of human experiences. It favors frequently occurring categories, such as joy, surprise, and no emotion. In comparison, the average scores for the emotions evoked given entire chains (columns labeled C) generated by the three methods is .24, .19 and .27 respectively. This is considerably higher than the average probability of the emotion on events (.08). The standard deviation shows substantial variability which indicates that there are several cases where the intended emotion has a very high probability. The generated context does disambiguate emotion categorization for our automatic system.

The chains generated through PCR evoke the desired emotions more effectively than those produced with the baseline method, as confirmed by human annotations (see Figure 2). PCR is particularly successful in generating event chains for

		Baseline			PC		PCR	
		E	B	C	B	C	B	C
Emotion-Specific Subsets	Anger	.02	.36	.27	.25	.17 −.10	.32	.24 −.03
	Boredom	.01	.09	.07	.08	.07 ±.00	.14	.12 +.05
	Disgust	.00	.22	.16	.25	.17 +.01	.30	.24 +.08
	Fear	.01	.46	.28	.30	.15 −.13	.53	.28 ±.00
	Guilt	.00	.31	.24	.24	.16 −.08	.34	.24 ±.00
	Joy	.33	.11	.24	.12	.26 +.02	.15	.27 +.03
	Pride	.01	.23	.23	.16	.15 −.08	.35	.24 +.01
	Relief	.08	.33	.56	.35	.53 −.03	.32	.61 +.05
	Sadness	.01	.37	.31	.32	.23 −.08	.50	.42 +.11
	Shame	.00	.28	.23	.20	.16 −.07	.32	.27 +.04
	Surprise	.14	.19	.37	.14	.24 −.13	.15	.30 −.07
	Trust	.02	.08	.07	.11	.08 +.01	.19	.13 +.06
	No-Emot.	.37	.02	.06	.03	.06 ±.00	.04	.09 +.03
Overall		.08	.23	.24	.20	.19 −.05	.28	.27 +.03
	σ	.24	.38	.37	.36	.34	.41	.38

Table 6: Averaged results for probabilistic emotion analysis given events (E) or, as generated by different methods for prompted emotions, backstories (B) or entire event chains (C). Delta values (marked with +/±/−) show the difference in comparison to the Baseline.

the emotions of relief and sadness. However, capturing emotions such as boredom or trust proves to be challenging, possibly due to the insufficient information in the context or backstory, or because these emotions require very specific events to be effectively evoked. Joy appears to be less distinctly recognized in contextualized events compared to isolated ones, suggesting that this emotion category does not benefit from our approach. Further analysis of discrete predicted labels reveals common misclassifications, particularly between guilt and shame, as well as between disgust and anger (see Appendix G). Moving forward, generating a larger initial sample size of target events, along with a final filtering process for unsuccessful cases, could produce a large number of narratives where the intended emotion responses are evoked for these particular categories.

In summary, our analysis shows that systems recognize modifications to the emotion interpretation of events under the inclusion of contextual backstories. Nonetheless, certain categories present challenges for clear emotion evocation through contexts, necessitating further exploration.

Investigating the Specific Influence of Backstories in Emotion Analysis. We now aim to understand how the generated backstories facilitate the observed modifications. Therefore, we investigate the emotion analysis derived solely based on the backstories, i.e. event chains without the target event (see Table 6, under the columns labeled B). We compare these values to the values for entire

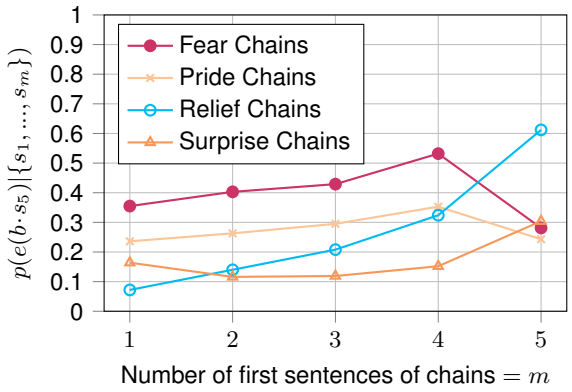

Figure 3: Mean of trajectories of predicted emotion probability scores in event chains for selected emotions ($\{s_1, ..., s_5\}$, generation method: PCR).

chains (columns labeled C) to measure how specifically the incorporation of the target event alters the emotion interpretation.

For certain emotions, such as relief and surprise, the PCR backstories alone show a substantially lower probability of evoking the specific emotion compared to the entire event chain. In contrast, for emotions like fear and pride, the PCR backstories have a higher probability of eliciting the intended emotion than the complete chain. These distinct patterns show that the manner of emotion evocation through generated backstories varies across different categories. While some emotions can be effectively elicited through the contextualization of the target event, others are predominantly grounded in the narrative of the backstory.

To explore this further, we investigate the trajectories of emotion prediction throughout the narratives, analyzing how the emotion scores evolve from sentence to sentence (see Figure 3). These plots not only show the probabilities associated with backstories (represented by $m = 4$ in the plots) and entire event chains ($m = 5$) but also probabilities derived from preceding segments of the narrative chain (indicated by $m \leq 3$ values). The plots illustrate the emotion trajectory for the four prominent categories previously discussed. More detailed visualizations of all emotions are given in Appendix H.

For the emotions fear and pride, we observe a steady increase followed by a decline in probability upon the introduction of the target event. This decline can be explained by our earlier findings, which suggest that the generated events (s_5) are not particularly aligned with these emotions, indicating that these have to derive from the context established in the backstory.[6] Conversely, the emotion relief exhibits a steady increase in probability throughout the narrative, with the most significant surge occurring upon the introduction of the final event. This pattern aligns with Ekman's concept of "programs" (as further described by Troiano et al., 2023), as such complex emotions develop through a progression, which is realized in our narratives. Table 7 shows this on examples generated by PCR for the category relief: the five stories with the highest emotion delta, i.e. score increase from backstory to the full chain in emotion analysis.

Our analysis shows that generated backstories can influence the emotion analysis of events in different ways. It is important to note that in certain cases, evoking emotions mainly from the backstory is not an inherent drawback. In fact, this approach can be essential for eliciting complex emotions that would otherwise remain unexplored. Our data contain events with ambiguous emotions as well as events devoid of specific emotion, which could potentially be emotionally charged through contextualization. An investigation of our data shows that, if the model shows low uncertainty in emotion analysis for events, this indicates that these events are easier to disambiguate using coherent narratives (Appendix I). Further distinction between such types of events justifies additional analysis and is a starting point for future work.

In summary, our analyses underscore the complex interplay between context and emotion evocation, highlighting that the foundation established by a narrative backstory can be critical for the perception and experience of particular emotions, while also revealing distinctive pathways for how emotions manifest within narratives.

5. Conclusion

We explore the relationship between contextual backstories and emotion interpretations in event descriptions, in cases in which events alone do not provide sufficient context for a non-ambiguous interpretation. We therefore make potential interpretations by readers explicit. Our paper contributes a dataset to study such additional context and proposes a set of methods, inspired by story planning, to achieve coherent and diverse narratives. Indeed, our additional context does increase inter-annotator agreement and is correctly recognized by automatic prediction systems. Specifically, our approach substantially enhances the evocation of emotions such as relief and sadness, while emotions like boredom and trust are less effectively evoked with the additional context we provide. A more comprehensive analysis using further emotion-labeled story datasets is a valuable future task to enhance our findings. Concerning the

[6] We chose to retain these instances in our analysis to evaluate the effectiveness of our method across any events and emotions. However, for future applications of the dataset, it may be advisable to filter these cases.

I lost my grandfather's calculator, a gift I had treasured for years. The teacher announced a crucial math problems exercise, one that would determine our grades for the semester. I rummaged through my backpack and desk, but couldn't find a spare calculator. The teacher began explaining the exercise, and I realized I had no way to participate without a calculator. *The teacher passed around a calculator for us to use during the math problems exercise.*
I received an urgent work assignment with a tight deadline that required my undivided attention. My roommate threw an impromptu party in the living room, immediately filling the air with loud music and chatter. The music grew louder, making it increasingly difficult for me to concentrate on my task. The party reached its peak, with the music blasting at an ear-splitting decibel that made it impossible for me to focus. *The music suddenly stopped when someone accidentally knocked over the bluetooth speaker.*
My friend pulled out his pocket knife and started flipping it open and closed near the campfire, narrowly missing his own fingers. As he continued to play with the knife, he began to twirl it around his fingers, coming close to slashing his own arm. I asked him to stop, but he ignored me, using the knife to cut a branch that snapped back and almost hit me. The tension between us grew as he continued to handle the knife carelessly, causing me to flinch every time he came near me with it. *As we were setting up camp near the lake, my friend accidentally dropped his pocket knife into the water.*
I stopped to take a photo of the waterfall and lost sight of the group in the dense foliage. I walked a short distance down the trail, expecting to catch up with the group, but they were nowhere in sight. As I turned back to retrace my steps, I realized I had taken a wrong turn and was now facing an unfamiliar section of the trail. The sound of the waterfall grew fainter, and I was surrounded by an unsettling silence, with no sign of the group anywhere. *Our guide blew the rescue whistle to signal for us to regroup after we got separated while hiking near the waterfall.*
I got lost in the city while trying to find a restaurant for lunch, unable to read the signs or ask for help due to the language barrier. I tried to ask multiple locals for directions, but none spoke English, and I was forced to rely on hand gestures and guesswork. I finally found my way back to the hotel, late for a group meeting, where my fellow travelers were worried about me and struggled to communicate with the hotel staff themselves. I received a reminder about my upcoming meeting with a local business partner, who I knew didn't speak English, and I began to worry about how I would negotiate a crucial business deal. *The tour guide handed out translation devices to help us navigate the language barrier during our trip abroad.*

Table 7: Examples of relief stories with highest emotion delta (score increase from backstory to full chain).

language-dependence of our work, Schäfer et al. (2025) have demonstrated similar results for German narratives; however, their findings suggest that the generated data is less effective in eliciting desired emotions in that language. This discrepancy likely arises because the large language model is more proficient in English.

Next to the modeling and dataset perspective, our method can also be seen as a contribution to Explainable Artificial Intelligence (XAI). By generating contextual backstories, our approach can be used to clarify uncertainties and probabilities in automatic classifications. This allows model decisions to be communicated more clearly and alternative interpretations to be demonstrated.

Our dataset EBS (Emotional BackStories) can be used to explore various dimensions, such as emotion changes in event chains or the impact of demographic differences on contextualized emotion interpretations. For example, Schäfer et al. (2026) show how it can be utilized to examine how appraisals evolve in the context of emotions, particularly by investigating the trajectories of appraisal scores within narratives. Consequently, the dataset offers further understanding of how emotions can be modeled from textual event descriptions. Furthermore, our dataset can be instrumental in distinguishing between types of events where initially the evoked emotions are unclear – specifically, those that are devoid of emotion versus those that evoke multiple emotions. This distinction underscores how contextual backstories can influence emotion responses in particular events and contributes to clarifying scenarios where initial emotion analysis may be unclear. In our dataset, variations in how backstories modify emotions elicited by events reveal distinct patterns, which can help differentiate between these types of events. Moreover, our dataset serves as novel training data for models specializing in contextualized emotion analysis. By encompassing a variety of backstories that elicit different emotion responses from the same event, researchers can use our data to develop and test nuanced classifiers.

While we provide valuable insights and a novel dataset, further research is needed to expand on our findings. Specifically, future work should investigate how our approach of generating narrative backstories and the resulting dataset can be tailored for various applications, such as improving dialogue systems or enhancing the recognition of human emotions in interactions with language models. In particular, the potential utility of our approach needs to be further explored in order to develop systems that better interpret the complexity of human emotions in different contexts.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) in the project KL2869/1–2 (CEAT, project number 380093645).

Limitations

While our study offers valuable insights into emotion analysis through generated narratives, it is important to acknowledge several limitations. First, our dataset is based on only 1,000 event descriptions generated by one model, which may limit the diversity and generalizability of our findings across different contexts and emotion categories. Incorporating a second, distinct language model to validate the emotion in the generated narratives could enhance robustness by discarding instances with mismatched emotions. Employing multiple language models in our analysis would further strengthen the significance of our experimental results. Second, the human annotation study was conducted solely on a sample of the data, which may not fully capture the complexity of the emotion responses present in the entire dataset. Additionally, the automatic model used to assess the entire corpus is the same one that generated the data, potentially reinforcing interpretations. Moreover, this model has not been validated for contextualized emotion analysis, leaving its accuracy uncertain in this context. To facilitate reproducibility, we offer access to our code, dataset, model predictions, and the human annotations, enabling further exploration and validation of our findings.

Ethical Considerations

We relied on crowd workers from Prolific to conduct the human evaluation. The annotators were paid £9.85–£10.59 per hour. All participants were shown a consent form containing the information and requirements regarding the study. They had to confirm their acceptance to be able to participate in the study. We provided an email address to contact us in case of problems during and after the study. The total cost of the annotation study including platform fees and taxes amounted to £511.64.

In developing our emotion analysis framework, we are committed to ethical considerations surrounding the use of generated narratives. We recognize the potential for biases in emotion interpretation, which may inadvertently reinforce stereotypes or misrepresent certain emotion contexts. To mitigate these risks, we emphasize transparency in our methodology and provide access to our datasets and models, allowing others to examine and address ethical implications in their applications. Additionally, we advocate for thorough evaluation of the generated narratives to ensure they align with ethical standards and contribute positively to understanding human emotions. Our work poses a risk of automatic emotion manipulation through the intentional evocation of specific feelings and may also reinforce biases in emotion classification due to the repeated use of automatic systems.

ChatGPT (OpenAI, 2025) was used to gain inspiration for formulations of initial notes for some of the text of this paper, as well as to find typos.

6. Bibliographical References

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Lynda Clark and Divij Sood. 2022. Working backwards: Creating a character backstory generation system using idealized creative writing outputs: Creating a character backstory generation system using idealized creative writing outputs. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, FDG '22, New York, NY, USA. Association for Computing Machinery.

Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. 2023. Emotion recognition based on psychological components in guided narratives for emotion regulation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 72–81, Dubrovnik, Croatia. Association for Computational Linguistics.

Dipankar Das, Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. 2011. Temporal analysis of sentiment events – a visual realization and tracking. In *Computational Linguistics and Intelligent Text Processing*, pages 417–428, Berlin, Heidelberg. Springer.

Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. 2023. Deep generative models for synthetic data: A survey. *IEEE Access*, 11:47304–47320.

Paul Ekman. 1972. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation 1971*, volume 19. Lincoln University of Nebraska Press.

Aline Etienne, Delphine Battistelli, and Gwénolé Lecorvé. 2022. A (psycho-)linguistically motivated scheme for annotating and exploring emotions in a genre-diverse corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 603–612, Marseille, France. European Language Resources Association.

Aline Étienne, Delphine Battistelli, and Gwénolé Lecorvé. 2024. Emotion identification for French in written texts: Considering modes of emotion expression as a step towards text complexity analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 168–185, Bangkok, Thailand. Association for Computational Linguistics.

Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024. A synthetic data approach for domain generalization of NLI models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2212–2226, Bangkok, Thailand. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052, Online. Association for Computational Linguistics.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Haein Jung, Heuiyeen Yeen, Jeehyun Lee, Minju Kim, Namo Bang, and Myoung-Wan Koo. 2023. Enhancing task-oriented dialog system with subjective knowledge: A large language model-based data augmentation framework. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 150–165, Prague, Czech Republic. Association for Computational Linguistics.

Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski, and Jan Kocoń. 2023. Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion*, 94:43–65.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. 2021. Can transformer models measure coherence in text: Re-thinking the shuffle test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1058–1064, Online. Association for Computational Linguistics.

Sofie Labat, Thomas Demeester, and Véronique Hoste. 2024. Emotwics: a corpus for modelling emotion trajectories in dutch customer service dialogues on twitter. *Language Resources and Evaluation*, 58(2):505–546.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Congda Ma, Kotaro Funakoshi, Kiyoaki Shirai, and Manabu Okumura. 2023. Coherent story generation with structured knowledge. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 681–690, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Meta. 2024. Llama (model llama-3.1-70b-instruct). Large language model.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, and Marcin Gruza. 2023. Capturing human perspectives in NLP: Questionnaires, annotations, and biases. In *Proceedings of the 2nd Workshop*

on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland.

Piotr Milkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocon. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259, Online. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David Chan. 2024. Virtual personas for language models via an anthology of backstories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19864–19897, Miami, Florida, USA. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

OpenAI. 2025. ChatGPT (model GPT4o-mini). Large language model.

Seo-Hui Park, Byung-Chull Bae, and Yun-Gyung Cheong. 2020. Emotion recognition from text stories using an emotion embedding model. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 579–583.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(4):344–350.

Evgeniia Razumovskaia, Joshua Maynez, Annie Louis, Mirella Lapata, and Shashi Narayan. 2024. Little red riding hood goes around the globe: Crosslingual story planning and generation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10616–10631, Torino, Italia. ELRA and ICCL.

Johannes Schäfer, Janne Wagner, and Roman Klinger. 2026. Appraisal trajectories in narratives reveal distinct patterns of emotion evocation. In *Proceedings of the 15th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Rabat, Marocco. Association for Computational Linguistics.

Johannes Schäfer, Sabine Weber, and Roman Klinger. 2025. Localization of English affective narrative generation to German. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 241–256, Hannover, Germany. HsH Applied Academics.

Armand Stricker and Patrick Paroubek. 2024. Chitchat as interference: Adding user backstories to task-oriented dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3203–3214, Torino, Italia. ELRA and ICCL.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Haruya Suzuki, Sora Tarumoto, Tomoyuki Kajiwara, Takashi Ninomiya, Yuta Nakashima, and Hajime Nagahara. 2022. Emotional intensity estimation based on writer's personality. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 1–7, Online. Association for Computational Linguistics.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Eileen Wemmer, Sofie Labat, and Roman Klinger. 2024. EmoProgress: Cumulated emotion progression analysis in dreams and customer service dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5660–5677, Torino, Italia. ELRA and ICCL.

Kaige Xie and Mark Riedl. 2024. Creating suspenseful stories: Iterative planning with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2391–2407, St. Julian's, Malta. Association for Computational Linguistics.

Dingyi Yang and Qin Jin. 2024. What makes a good story and how can we measure it? a comprehensive survey of story evaluation.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A. Event Types and Objects

Table 8 shows our list of 10 event types along with examples of 20 corresponding objects. These values serve as attributes in the prompt which initially generates diverse events that we subsequently develop backstories for.

B. Full Text Prompts

Table 9 shows the full text of the prompts we use to generate our data according to the different methods. The interaction with the model comprises three message types[7]: a "system" message that establishes the context for the interaction and includes general guidelines; a "user" message that encapsulates the specific inputs, requirements, and instructions for the task; and an "assistant" message that represents the model's response based on the provided context.

Prompt 1 is used to generate the concluding event and aims to ensure a balanced topic variance through the incorporation of specific attributes. To ensure that the generated event descriptions are concise, we implement few-shot prompting by including ten examples of event descriptions in the prompt. As values for the different attributes, we use 10 distinct event types, each associated with 100 unique objects. We acquire these values from several interactions with ChatGPT (OpenAI, 2025). For example, an event type can be a "Social Gathering" with objects like "balloons". The full list of event types with examples of objects is given in Appendix A.

For each event generated in the first step, we aim to create a corresponding backstory comprised of four preceding events, tailored to influence the specific emotion responses of the last event. In the baseline approach, this process is implemented using a single prompt (**Prompt 2** as shown in Table 9). For the other methods, the process is further broken down into sub-steps: Story planning using **Prompt 2.1:** For each final event, the LLM is prompted to first generate a story plan that outlines plausible explanations for the emotion responses tied to the concluding event. This preemptive planning increases coherence and relevance in the backstories. Based on the crafted story plan, the LLM should generate the sequential backstory comprising four events. This ensures that the events are narratively connected and describe the context leading up to the pivotal final event. Event chain generation using **Prompt 2.2:** After generating the event chain, we conduct a summarization step to ensure uniformity and clarity. Additionally, our third method uses **Prompt 2.3**

[7] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/

Event Type	Event Type Objects
Social Gathering	chairs, tables, food platters, drinks, napkins, decorations, music speakers, games, invitation cards, host, guests, tablecloths, candles, balloons, party favors, photo booth, name tags, cutlery, glasses, ice bucket, ...
Educational Activity	textbooks, notebooks, pencils, whiteboard, markers, projector, handouts, calculator, overhead projector, globe, poster board, computer, scissors, gluestick, craft supplies, timers, textual resources, reference books, tables, student desks, ...
Recreational and Nature Activity	hiking boots, backpacks, water bottles, first-aid kit, campfire supplies, nature guide, binoculars, tent, sleeping bags, camping chairs, fishing gear, bicycles, kayaks, picnic basket, coolers, maps, sunscreen, bug spray, fishing rods, swimming gear, ...
Cultural and Community Event	stage, performers, sound system, projector, festival tickets, food stalls, craft booths, cultural displays, artworks, costumes, brochures, community posters, instruments, banners, seating areas, local products, vendors, volunteers, refreshments, cultural symbols, ...
Professional Development	business cards, presentation slides, notebooks, pens, projector, handouts, networking tools, feedback forms, laptops, name badges, workshops, career fair flyers, industry reports, coffee cups, panel discussion guides, training materials, lecture notes, team-building activities, case studies, clipboards, ...
Celebration	cake, candles, party hats, balloons, confetti, party favors, streamers, drinks, gift bags, music playlist, photo booth, decorations, invitation cards, celebration banner, tables, chairs, food platters, glasses, plates, silverware, ...
Artistic Performance	stage, costumes, sets, props, lights, sound equipment, musical instruments, audience seats, backdrops, tickets, makeup kit, rehearsal schedule, choreography notes, great hits collection, piano, amplifiers, performance schedule, music sheets, playbill, actors, ...
Competition	trophies, medals, referee kit, scoreboard, team jerseys, game equipment, whistle, competition schedule, event tickets, player registration, crowd barriers, timing devices, venue maps, registration forms, team banners, score sheets, first aid kits, video cameras, performance analytics, heat sheets, ...
Family and Relationships	family photo albums, toys, family tree chart, gift cards, family recipe book, family calendars, cameraman, outdoor equipment, gifts, personalized items, family game night materials, storybooks, name tags, family bonding games, sentimental objects, blankets, picnic spreads, board games, interactive toys, family outings, ...
Transportation and Travel Event	maps, itineraries, backpacks, suitcases, boarding passes, tickets, travel guides, snacks, passports, travel pillows, sunscreen, water bottles, portable chargers, cameras, compact umbrellas, guidebooks, tour buses, airport shuttles, magazine subscriptions, reservation forms, ...

Table 8: List of event types and objects used in prompt templates to create diverse event descriptions.

to revise the generated event chain in an attempt to ensure adherence to the defined requirements.

C. Descriptive Dataset Statistics

To assess lexical diversity, we analyze the unigrams used in the backstories as well as Jaccard-based diversity scores. Additionally, we analyze the coherence of event chains.

We perform tokenization for each event description by removing punctuation symbols, converting all text to lowercase, and splitting the resulting string on whitespace characters. Event descriptions of the final events s_5 in our dataset consist of 15.4 tokens on average. The length of the event descriptions in the backstories are shown in Table 10.

C.1. Analysis of Lexical Diversity

For each of the three data generation methods, we identify the most frequent unigrams (nouns) overall and among emotion-specific subsets.

Additionally, we compute a comparative lexical diversity measure using Jaccard-based diversity scores, calculated through pairwise comparisons of the word types utilized in the backstories generated by the different methods. We define this diversity score between two instances b_i and b_j as $D(b_i, b_j) = 1 - J(b_i, b_j)$. The Jaccard coefficient (Jaccard, 1912) for these instances is defined as

$$J(b_i, b_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|}$$

where T_i and T_j represent the sets of unique word types in instances b_i and b_j, respectively. With this we define the diversity in a set of instances, e.g. a subset of backstories $B = \{b_1, \ldots, b_n\}$, as

$$D(B) = 1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} J(b_i, b_j)}{n^2}.$$

A high diversity score indicates that the backstories are highly different from each other, reflecting greater lexical diversity, while a low score suggests that the backstories share many common word types and are thus more similar.

C.2. Coherence Scoring Algorithm

To assess the coherence of event chains, we adopt a zero-shot shuffle test methodology as suggested by Laban et al. (2021). In the shuffle test, the texts are divided into smaller units, such as sentences, which are subsequently shuffled to create permutations. A large language model is employed to evaluate the coherence of the original and each permutation by extracting the likelihood predictions for the token sequences. Lower likelihood scores are interpreted to be indicative of permutations that exhibit reduced coherence. The framework proposed by Laban et al. (2021) highlights that employing a zero-shot setting, wherein the language model is utilized without prior fine-tuning specific to the shuffle test, facilitates a more accurate assessment of coherence.

We implement this methodology to compute coherence scores for each event chain $c =$

#	MT	Prompt Text
Prompt 1	system	You are a person describing an event which you have experienced. 10 examples of such event descriptions are as follows: 0: The phone rang. 1: A cat meowed. 2: The car engine sputtered to a stop. 3: A child laughed in the park. 4: A bird fluttered past the window. 5: The waves crashed against the shore. 6: A train whistled as it approached. 7: The fireworks lit up the sky. 8: A bicycle rode by. 9: A crowd cheered at the concert.
	user	The event you experienced is of type: {ds_event_type}. In a longer text you are describing several things that happened at that event. Something happened at that event with the following object(s): {ds_event_object}. In your response, only provide a very short sentence describing what happened to/with the object(s).
Prompt 2	system	It is often clear from the text that describes an event which specific emotion it evokes in a person that experienced it. However, additional information about the situation can change our understanding of how a person might interpret the event. You are an expert at creating a scenario that explains why a specific event may cause a possibly unusual emotion in you. In addition, you can concisely make this scenario apparent for the reader by formulating a description of 4 events that took place immediately before the event.
	user	You experienced something happening which is described by the following event description: 5. "{event}". This event somehow made you clearly feel the emotion: "{emotion}". Provide a text describing four events that took place immediately before event 5 by giving a list of descriptions of these events (1.-4.). The events 1.-4. clearly influence your personal emotional interpretation of the event that happened after (5.). The emotion "{emotion}" is only triggered by what specifically happened in event 5. The events 1.-4. evoked other emotions, such as: {", ".join(random.sample([emo for emo in EMOTION_SET if emo != emotion], len(EMOTION_SET)-1))}. In your response, for each of the 4 event descriptions: Only give a summary text consisting of the main clause in a very short sentence. Each description should only describe a singular event. Indicate each event description in a separate line.
Prompt 2.1	system	It is often clear from the text that describes an event which specific emotion it evokes in a person that experienced it. However, additional information about the situation can change our understanding of how a person might interpret the event. You are an expert at creating a scenario that explains why a specific event may cause a possibly unusual emotion in you. In addition, you can concisely make this scenario apparent for the reader by formulating a description of 4 events that took place immediately before the event.
	user	You experienced something happening which is described by the following event description: 5. "event". This event somehow made you clearly feel the emotion: "emotion". First, give a brief explanation of a scenario in which it can be deduced from the description of event 5. that you felt emotion. Second, phrase this explanation as events that took place immediately before event 5 by giving a list of descriptions of these events (1.-4.). The events 1.-4. clearly influence your personal emotional interpretation of the event that happened after (5.). The emotion "emotion" is only triggered by what specifically happened in event 5. The events 1.-4. evoked other emotions, such as: {", ".join(random.sample([emo for emo in EMOTION_SET if emo $\neq$ emotion], len(EMOTION_SET)-1))}.
Pr. 2.2	user	Extract the sequence of 4 descriptions of events that happened from the following text: ### {explanation} ### The event 5: "{event}" happened after the 4 events. In your response, for each of the 4 event descriptions: Only give a summary text consisting of the main clause in a very short sentence. Each description should only describe a singular event. Indicate each event description in a separate line.
Prompt 2.3	system	You are an expert at adapting a narrative to convey specific emotional interpretations. You will receive a text that outlines a sequence of events as experienced by an individual. Additionally, there will be an explanation of how a particular emotion is triggered in this individual based on the final event.
	user	Explanation: {story_plan} Event sequence: {chain} First, provide a brief evaluation on how the first four events (1.-4.) of the sequence could be adjusted to form a coherent narrative which better aligns with the conclusion given in the explanation. The text of the last event (5.) should remain as is. Second, provide a revised event sequence that incorporates these adjustments while keeping the sentence length for each event description similar. Each event description should consist only of a main clause in a very short sentence. Do not explicitly mention the emotions felt.

Table 9: List of prompts used in our methods for event chain generation. The first column shows the numbers of the specific prompts as introduced in Figure 1. 'MT' refers to the prompt message type as specified in the input for the instruction-tuned LLM.

Method	s_1	s_2	s_3	s_4	Σ
Baseline	16.4	16.5	17.6	19.4	69.8
PC	12.2	12.1	12.2	12.7	49.2
PCR	18.3	19.0	20.2	21.6	79.2

Table 10: Text length in number of tokens of event descriptions in backstories generated by different methods comprising four sentences each.

$(s_1, \ldots, s_l)$ within our dataset, where $l = 5$. The adapted algorithm is detailed as follows.

1. Shuffling. We include shuffled variants for each event chain c by computing the set of its permutations $\mathfrak{S}_c = \{\bar{c}_1, \ldots, \bar{c}_k\}$, where $\bar{c}_1 = c$ and $k = l!$ (in our dataset, $k = 5! = 120$). To reduce the computational cost, we randomly sample $\frac{k}{4} = 30$ permutations from this set, resulting in $\mathfrak{S}'_c$, while ensuring that c is included in this sample.

2. Language Modeling. We encode the original event chain and the sampled shuffled variants using a language model M[8]. M implements a function g which tokenizes[9] each $\bar{c} \in \mathfrak{S}'_c$ into a sequence of tokens $g(\bar{c}) = (t_1, \ldots, t_n)$. For each tokenized sequence $g(\bar{c})$, M computes a sequence of prediction score vectors as output of its language modeling head, i.e. the logits of M, as $f(g(\bar{c})) = f((t_1, \ldots, t_n)) = (l_1, \ldots, l_n)$. We transform these scores into values which we can interpret as log probabilities of transitions by applying a softmax followed by a logarithm for each token position as

$$h_i = \log(\text{Softmax}(l_i)) = [\log(\frac{e^{l_{ij}}}{\sum_{q=1}^{v} e^{l_{iq}}})]$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, v$, where v is the size of the vocabulary of M. The log-likelihood of each sequence is computed based on the transition probabilities as

$$\log P(\bar{c}) = \sum_{i=1}^{n} \log P(t_i | t_0 \ldots t_{i-1}) = \sum_{i=1}^{n} h_{i,w(t_{i+1})}$$

where $w(t_{i+1})$ is the index of $t_{i+1} \in g(\bar{c})$ in the vocabulary of M and $h_{i,w}$ denotes the value in position w of the vector h_i.

For the original event chain and each of its shuffled permutations $\bar{c} \in \mathfrak{S}'_c$, we calculate their log-likelihood scores $\log P(\bar{c})$.

[8]We utilize the Llama-3.1-8B-Instruct model from the Meta-Llama collection, accessible via HuggingFace: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

[9]Before tokenization, the texts of the events in each event sequence are combined into a single string using a whitespace character as separator.

3. Coherence Score Calculation. We continue by ranking the original chain's likelihood against those of all its shuffled permutations. The coherence score for the original chain is computed as

$$H(c) = 1 - \frac{r(c)}{|\mathfrak{S}'_c|}$$

where $r(c)$ is the rank of the original log-probability $\log P(c)$ among the permutation scores, calculated as

$$r(c) = |\{\bar{c} \in \mathfrak{S}'_c \mid \log P(\bar{c}) \geq \log P(c)\}| \,.$$

A higher score $H(c)$, approaching 1, indicates a high coherence of the event chain, while a score closer to 0 reflects low coherence.

C.3. Results

This section presents additional statistical analyses of our dataset, which consists of 13,000 event chains for each generation method. Each event chain comprises five short texts, each a single sentence representing event descriptions from the perspective of an event experiencer, organized in chronological order. The dataset is structured around 1,000 pivotal events, each enriched by four preceding events, collectively referred to as backstories, as detailed in Section 3.2. Each pivotal event is paired with a unique backstory corresponding to one emotion, as outlined in Section 3.1, resulting in a comprehensive dataset that facilitates emotion analysis across various contexts.

Unigrams. The top 20 unigrams in the backstories generated by the different methods are shown in Table 11.

Additionally, Table 12 shows unique items within the top 100 unigrams pertinent to each emotion category for each method. This table reveals further distinctive characteristics tied to specific emotional narratives. For instance, in the anger category, "changes" may suggest conflicts or disruptions, while a word such as "wedding" in the sadness category evokes nostalgic themes and significant life events. Additionally, the term "proposal" in the pride category suggests achievement and recognition, contrasting sharply with "trash" and "scandal" in the disgust category, which imply negative experiences. While some terms may appear less characteristic of the emotions, several others resonate clearly with the themes we expect to find in stories related to the corresponding emotions.

Lexical Diversity based on Jaccard Coefficient. We further analyze lexical diversity using average pairwise Jaccard-based diversity scores, summarized in Table 13. For clarity, the Jaccard coefficient allows us to yield diversity scores that reflect the degree of dissimilarity between backstories. A high

Subset	Baseline	PC	PCR
anger	event, hours, organizer, friend, day, team, conference, months, party, morning, call, project, family, group, staff, presentation, organizers, sister, hour, front	event, team, organizers, organizer, staff, group, friend, family, morning, email, call, hours, venue, day, instructor, meeting, conference, project, member, party	event, organizers, team, hours, family, staff, organizer, group, friend, concerns, time, venue, day, project, email, conference, party, teams, room, stage
boredom	friend, hours, event, hour, day, presentation, minutes, call, conference, project, argument, organizer, morning, party, group, time, team, coffee, family, front	friend, event, team, project, group, argument, organizer, call, coffee, family, instructor, conference, presentation, hours, venue, morning, stage, organizers, **train**, staff	event, time, friend, team, hours, stage, presentation, family, group, day, **crowd**, organizers, project, teams, conference, room, organizer, argument, venue, morning
disgust	event, friend, organizer, hours, group, **food**, conference, **company**, day, morning, hour, staff, party, family, organizers, project, team, room, speech, front	event, organizer, friend, staff, organizers, group, team, **food**, family, conversation, room, member, venue, conference, stage, **companys**, call, meeting, morning, instructor	event, organizers, **food**, staff, organizer, friend, group, family, team, concerns, room, **hands**, conference, equipment, **trash**, member, **companys**, stage, venue, **company**
fear	event, call, friend, day, **warning**, family, organizer, conference, morning, hours, group, email, **fire**, project, **phone**, team, **area**, **message**, **performance**, presentation	call, friend, group, event, team, family, email, conversation, project, **message**, colleague, stage, room, meeting, organizer, argument, venue, presentation, morning, organizers	family, event, team, friend, group, room, feeling, project, **eyes**, stage, call, email, **message**, **mind**, time, **performance**, presentation, teams, **safety**, venue
guilt	friend, event, day, sister, organizer, family, team, morning, party, project, call, group, conference, time, **colleague**, months, hours, work, night, presentation	friend, event, team, call, colleague, family, project, friends, morning, email, organizer, group, conversation, party, argument, meeting, organizers, day, concerns, room	friend, event, team, friends, family, project, colleague, time, concerns, group, call, party, organizers, day, feeling, organizer, morning, **task**, attendees, work
joy	friend, hours, event, day, family, call, morning, sister, argument, months, project, team, conference, time, email, hour, party, organizer, presentation, coffee	team, friend, family, event, hours, group, venue, email, morning, call, project, argument, conversation, stage, day, sound, **teams**, party, room, time	team, event, family, friend, time, hours, venue, room, day, teams, stage, party, project, group, sound, morning, friends, email, call, issues
pride	event, team, hours, morning, project, months, family, night, day, time, organizer, friend, group, conference, teams, party, presentation, sister, work, organizers	team, event, hours, call, group, email, friend, concerns, morning, venue, **system**, family, meeting, **night**, equipment, member, organizers, argument, **design**, conference	team, event, hours, concerns, time, teams, **system**, group, family, **design**, issue, **plan**, work, project, friend, venue, equipment, attendees, meeting, party
relief	hours, event, day, call, organizer, time, team, morning, argument, hour, friend, minutes, conference, presentation, night, family, party, group, sister, front	team, call, event, friend, group, hours, family, room, stage, member, argument, time, morning, staff, sound, equipment, project, instructor, venue, meeting	team, time, event, hours, room, family, stage, venue, equipment, group, call, sound, friend, attendees, concerns, schedule, teams, argument, project, **solution**
sadness	friend, call, family, sister, event, **grandmother**, day, morning, months, **years**, hours, party, conference, team, organizer, project, argument, **concert**, time, **grandmothers**	friend, family, call, event, **grandmothers**, **grandmother**, venue, morning, team, group, friends, email, conversation, **photo**, **grandfathers**, project, party, day, **sister**, organizers	friend, **grandmother**, family, **grandmothers**, event, **memories**, **grandfather**, friends, time, **grandfathers**, team, call, venue, hours, **photo**, project, group, morning, day, stage
shame	event, friend, organizer, family, day, group, team, presentation, **friends**, morning, front, party, hours, months, **skills**, conference, project, organizers, speech, sister	friend, friends, team, event, group, family, call, colleague, organizer, project, party, email, morning, conference, presentation, day, **skills**, venue, organizers, argument	friend, event, friends, team, group, family, time, party, project, colleague, **skills**, morning, day, organizer, conference, organizers, everything, attendees, feeling, concerns
surprise	event, hours, organizer, friend, day, morning, conference, staff, party, minutes, hour, team, room, family, stage, argument, time, **issues**, presentation, organizers	event, team, friend, organizer, organizers, staff, group, venue, argument, instructor, room, stage, family, conference, email, member, conversation, call, morning, **attendees**	event, team, staff, organizers, organizer, room, friend, venue, stage, time, group, attendees, family, hours, conference, teams, equipment, concerns, instructor, sound
trust	friend, event, organizer, hours, staff, group, team, argument, family, sister, party, conference, day, call, **instructor**, presentation, teams, **speaker**, morning, stage	team, friend, event, staff, group, organizers, argument, instructor, venue, meeting, family, member, concerns, stage, **leader**, organizer, project, hours, **line**, conversation	team, staff, event, friend, organizers, concerns, group, family, stage, time, venue, member, instructor, issues, issue, meeting, teams, project, room, sound
no-emotion	event, friend, hours, day, call, hour, argument, family, organizer, conference, party, project, team, time, morning, work, minutes, coffee, **boss**, group	team, call, friend, event, argument, project, group, morning, email, **issue**, family, organizer, stage, colleague, time, equipment, coffee, member, venue, presentation	team, event, friend, argument, hours, time, issue, call, schedule, project, group, family, room, equipment, venue, stage, morning, everything, presentation, sound
Overall	event, friend, hours, day, organizer, family, morning, team, call, conference, party, group, project, sister, hour, presentation, months, argument, time, staff	team, friend, event, group, call, family, organizer, project, venue, email, morning, organizers, staff, argument, hours, instructor, conversation, stage, conference, meeting	event, team, friend, family, time, group, hours, organizers, room, project, staff, stage, venue, concerns, organizer, day, teams, friends, conference, call

Table 11: Top 20 unigrams (nouns only) in backstories generated by different methods. Unigrams marked in bold are unique for the set of chains of the respective emotion.

(or low) diversity score indicates that instances are very different from (or similar to) each other, reflecting correspondingly high (or low) diversity. Overall, this analysis indicates a consistent level of lexical diversity throughout the dataset. The overall lack of biases in lexical diversity across different emotion categories further underscores the homogeneity of the data, implying that the narratives were constructed as diverse as intended without influence from the emotion categorization.

D. Examples of Generated Event Chains

Table 14 shows narratives generated by our three methods for four emotion categories. On closer inspection, we can see that, compared to the other approaches, the Baseline tends to link the backstory very strongly to the last event and frequently reuse content mentioned in it. In contrast, the narratives of the other two methods seem to be more complex and go further into the background. The PCR method in particular conveys the narrative more clearly by emphasizing the backstory more

Subset	Baseline	PC	PCR
anger	—	changes, weekend, booth	changes
boredom	tray, guest, spot	performer, alarm, keynote, drinks, course, lecture, blew, opening, drink, minute	act, routine, week, break, keynote, alarm
disgust	trash, sandwich, impact, festivals, sponsor, field	hands, scandal, article, bag, practices, history, spill, document	hands, trash, smell, table, history, scandal, floor, tone, health, waste, practices, bag, signs, person, dirty, impact, sustainability
fear	message, security, number, venues, history, department, workshop, accident	security, classmate, sign, child, office	warning, something, security, job, expression
guilt	struggles, gift, responsibility, help, sisters, decision	task, sibling, care, budget	decision, responsibility, glimpse, disappointment, media, tasks, sibling, didnt
joy	—	pride	anticipation
pride	design, ability, proposal, setup, volunteer	ability, kids, idea, management, last-minute, encouragement	confidence, ability, proposal, action, determination, sponsor, abilities, program, management, guidance
relief	deadline, planner, backup	supplier, solution, delivery, traffic, situation, rehearsal, couldnt	supplier, traffic, frustration
sadness	grandmothers, grandfather, health, grandfathers, photo, father	grandmothers, photo, grandfathers, grandfather, mother, heartfelt, photos, belongings, invitation, house, school, injury, note, name, box, band	grandmothers, grandfather, grandfathers, photo, stories, life, photos, memory, years, mother, celebration, share, contrast, note, father, wedding, countless, heartfelt, days
shame	coach, art, classmates, colleagues, name, aunt	skills, everyone, attention, importance, progress, concern	expertise, classmates, someone, control
surprise	weather, forecast, crew	presenter, performers	arrival, crew, performers, doubts
trust	smile, stranger, kind, apology, passenger, mistakes	mistakes, process, entrance, vendor	entrance, questions, ticket
no-emotion	times, plans	home, nod, supervisor, stand, speakers	workshop, home, text, weight, emergency

Table 12: Unique words for each emotion category in the top 100 unigrams (nouns only) in backstories generated by different methods.

Subset	Baseline	PC	PCR
anger	.87	.90	.89
boredom	.87	.90	.90
disgust	.87	.90	.90
fear	.87	.89	.89
guilt	.87	.89	.89
joy	.86	.90	.90
pride	.87	.90	.90
relief	.87	.91	.90
sadness	.86	.89	.89
shame	.87	.89	.89
surprise	.87	.91	.91
trust	.88	.90	.90
no-emotion	.86	.89	.90
Overall	.87	.90	.90

Table 13: Average pairwise Jaccard-based diversity scores of backstories by different methods.

elaborately in the sequence of events. Overall, this illustrates how the content derived from the approach PC is adopted and enriched in various spots.

E. Human Annotation Setup and Analysis

We performed a human annotation of events and event chains with regard to emotion analysis and text quality. We use the platform Prolific[10] to find study participants. These are screened to be in the UK, have English as their first and native language, and have an approval rate of at least 95%.

[10] https://www.prolific.com/

E.1. Annotation of Events

We sample a set of 10 events (one from each event type category, cf. Appendix A) and collect 39 annotations for this sample.[11] Additionally, we sample 90 more events (nine from each event type category, cf. Appendix A) and collect 3 annotations each, so we can evaluate the quality of the generated texts on a sample of 100 events (10% of our dataset).

E.2. Annotation of Event Chains

We collect 780 annotations of event chains, which corresponds to three annotations per 13 backstories, generated by each of the two methods, for each of the 10 sampled events. Table 15 shows the annotated labels in comparison to the labels in the prompts with the baseline method. This annotation results in the evaluation scores as shown in Table 16. The comparison of the annotated labels to the labels in the prompts with the PCR method results in the evaluation scores as shown in Table 17. We also ask annotators to evaluate the quality of the narratives based on a five-point Likert scale regarding the contextual influence of the backstory, realism of the story and whether it is written by a human or by an AI. Results for this annotation (see Table 18) do not show clear differences between the two tested methods.

[11] We collect 39 annotations for each event since we intend to compare these to the annotations of 13 corresponding event chains (with 3 annotations each).

Em.	Baseline	PC	PCR
Guilt	I had been tasked with testing the loudspeaker system before the big event. My supervisor warned me that a malfunction would be disastrous for the company's reputation. I skipped the recommended final check to grab a quick lunch before the event started. The event host began the ceremony, and the loudspeaker was working perfectly, filling me with temporary relief. *The loudspeaker suddenly malfunctioned and went silent.*	I was tasked with setting up the sound system for a charity event. I noticed the previous technician had left a mess with wires and cables scattered all over the floor. I decided to skip some of the safety checks to meet the deadline. I received a call from my friend, reminding me how much the event meant to her. *The loudspeaker suddenly malfunctioned and went silent.*	A close friend entrusted me with setting up the sound system for a charity event. I faced a daunting task, as the previous technician had left the equipment in disarray. With time running out, I took a shortcut to meet the deadline, skipping some crucial safety checks. Just before the event started, my friend reminded me of its significance and the many people counting on its success. *The loudspeaker suddenly malfunctioned and went silent.*
Relief	The loudspeaker had been blasting earsplitting static for 10 minutes. I had just received an urgent phone call from the event organizer, warning me of a severe thunderstorm approaching. The previous speaker's presentation was a disaster, with him stumbling over his words and spilling water on his notes. I had spent the last hour setting up the outdoor stage, and a strong gust of wind had just blown over a music stand. *The loudspeaker suddenly malfunctioned and went silent.*	I was tasked with giving a presentation to a large crowd. The sound system malfunctioned, amplifying my voice to an ear-piercing level. The sound technician ignored the problem and chatted with someone. The audience covered their ears and looked at me with discomfort. *The loudspeaker suddenly malfunctioned and went silent.*	I stood before a sea of expectant faces, the weight of their attention crushing me. The sound system suddenly shrieked to life, my amplified voice piercing the air like a knife. The sound technician fumbled with the controls, oblivious to the cacophony they were creating. The audience winced in unison, their pained expressions mirroring the turmoil brewing inside me. *The loudspeaker suddenly malfunctioned and went silent.*
Fear	I had just given a speech to a large crowd, criticizing the company's management. The CEO, visibly angry, approached the microphone to respond to my accusations. The loudspeaker suddenly screeched, causing the audience to cover their ears in discomfort. The CEO's face turned red as he began to speak, but his voice was barely audible due to the screeching. *The loudspeaker suddenly malfunctioned and went silent.*	I arrived at a remote wilderness survival training camp. I learned the camp was located near a toxic waste site. A fellow trainee struggled with the physical demands and dropped out. I completed a challenging obstacle course. *The loudspeaker suddenly malfunctioned and went silent.*	I arrived at a remote wilderness survival training camp, where the instructors emphasized the importance of following loudspeaker instructions for safety. The instructors warned us about the toxic waste site nearby and explained that the loudspeaker would alert us to any changes in air quality. During the first exercise, I struggled to navigate the challenging terrain, but the loudspeaker provided crucial guidance, helping me stay on track. I completed a difficult obstacle course, relying heavily on the loudspeaker's instructions to avoid hazards and find the safest route. *The loudspeaker suddenly malfunctioned and went silent.*
Pride	I had just finished a long and difficult speech in front of a large audience. The loudspeaker had been malfunctioning throughout my presentation, causing me to struggle to be heard. The audience was getting restless and some people started to leave due to the poor sound quality. I had just made a crucial point that seemed to resonate with the remaining audience members. *The loudspeaker suddenly malfunctioned and went silent.*	I set up the sound system for the big conference. The event organizer informed me that the conference was running 30 minutes behind schedule. I listened to the same annoying music loop playing over and over again. The keynote speaker began to talk with the sound system working flawlessly. *The loudspeaker suddenly malfunctioned and went silent.*	I spent the entire morning upgrading the sound system with a new backup system to prevent technical issues. The event organizer informed me that the conference was running 30 minutes behind schedule, giving me extra time to test the new backup system. I used the extra time to run a series of tests on the sound system, trying to simulate potential failures. The keynote speaker began to talk, and the sound system was working flawlessly, but I was still waiting for a real test of the new backup system. *The loudspeaker suddenly malfunctioned and went silent.*

Table 14: Examples of event chains generated by different methods.

F. Correlation of Human and System Emotion Annotation

The correlations between human annotations and system predictions are summarized in Table 19. These findings are derived from the evaluation of 100 event annotations and 260 chain annotations (all with 3 annotators each). The correlations values for events exhibit greater fluctuations across categories, whereas those for chains are more consistent. The findings reveal that, for individual events, there are notable differences in the understanding of categories such as disgust, guilt, and shame between human annotators and the system, as evidenced by the low correlation in these categories. Conversely, for the chains, only the emotion category of boredom shows a disparity in recognition between human annotators and the system, indicated by similarly low correlation. The highest correlation coefficients for events are observed in the categories of joy, surprise, and no-emotion, while for chains, highest correlations are noted for anger, fear, pride, and sadness.

G. Predicted Emotion Analysis

We compare the emotion label which is predicted with the highest likelihood score by our automatic emotion analysis method $e(c)$ to the emotion label each event chain has been created with $e(b \cdot s_5)$. We evaluate this for each of our three data generation methods. Table 20 shows the results for the Baseline approach. Table 21 shows the results for our second approach, PC. Table 22 shows the results for our third approach, PCR.

$e(b \cdot s_5)$ \ e_A	Anger	Boredom	Disgust	Fear	Guilt	Joy	Pride	Relief	Sadness	Shame	Surprise	Trust	No-Emot.
Anger	10	0	4	0	1	0	0	3	2	1	1	1	7
Boredom	7	3	1	2	0	4	1	1	1	1	3	2	4
Disgust	9	0	2	0	2	1	0	1	3	0	5	0	7
Fear	4	0	1	9	2	1	1	2	4	0	1	0	5
Guilt	4	1	0	1	4	1	0	3	6	1	1	1	7
Joy	2	2	1	3	1	4	1	4	7	1	2	0	2
Pride	3	1	1	1	2	1	9	3	2	0	1	2	4
Relief	6	2	0	3	0	1	0	4	6	2	4	0	2
Sadness	3	2	1	2	1	1	1	1	7	0	3	3	5
Shame	1	2	2	1	3	0	3	3	2	7	1	0	5
Surprise	6	1	1	2	0	4	1	5	2	0	6	0	2
Trust	4	1	1	4	1	1	1	3	3	0	4	2	5
No-Emot.	4	1	0	2	1	0	0	4	9	1	2	0	6
Σ	63	16	15	30	18	19	18	37	54	14	34	11	61

Table 15: Counts of prompted emotion $e(b \cdot s_5)$ vs. annotated emotion e_A for event chains (generation method: Baseline).

	Anger	Boredom	Disgust	Fear	Guilt	Joy	Pride	Relief	Sadness	Shame	Surprise	Trust	No-Emot.	Ma.-Avg.
Precision	.33	.10	.07	.30	.13	.13	.30	.13	.23	.23	.20	.07	.20	.19
Recall	.16	.19	.13	.30	.22	.21	.50	.11	.13	.50	.18	.18	.10	.22
F1	.22	.13	.09	.30	.17	.16	.37	.12	.17	.32	.19	.10	.13	.19

Table 16: Scores for prompted emotion vs. annotated emotion for event chain (generation method: Baseline). Ma.-Avg.: Macro-Average.

	Anger	Boredom	Disgust	Fear	Guilt	Joy	Pride	Relief	Sadness	Shame	Surprise	Trust	No-Emot.	Ma.-Avg.
Precision	.17	.00	.20	.40	.20	.07	.33	.20	.50	.37	.20	.07	.17	.22
Recall	.12	.00	.55	.34	.50	.11	.27	.13	.24	.50	.20	.20	.08	.25
F1	.14	.00	.29	.37	.29	.08	.30	.16	.33	.42	.20	.10	.11	.21

Table 17: Scores for prompted emotion vs. annotated emotion for event chain (generation method: PCR). Ma.-Avg.: Macro-Average.

	Infl.			Written By	
	Yes	No	Real.	Human	AI
Baseline	249	141	3.56	2.98	3.23
PCR	264	126	3.52	2.95	3.21

Table 18: Human evaluation of the quality of the generated narratives: Binary values for the contextual influence of the backstory (Infl.); averaged five-point Likert scale values regarding the plausibility of the story (Real.) and whether it is written by a human or by an AI.

Emotion-Specific Annotation	Events	Chains
Anger	.46 ***	.44 ***
Boredom	.29 **	.15 *
Disgust	.14	.22 ***
Fear	.39 ***	.46 ***
Guilt	.10	.27 ***
Joy	.60 ***	.31 ***
Pride	.41 ***	.43 ***
Relief	.43 ***	.38 ***
Sadness	.30 **	.49 ***
Shame	.13	.38 ***
Surprise	.50 ***	.34 ***
Trust	.37 ***	.29 ***
No-Emotion	.55 ***	.22 ***
Overall ML Annot.	.47 ***	.37 ***

Table 19: Spearman correlation between human and automatic emotion annotation on events and chains. Significance levels are indicated as *: p < 0.05; **: p < 0.01; ***: p < 0.001. ML Annot.: multi-label annotation.

H. Emotion Trajectories in Event Chains

Based on our automatic emotion analysis, we evaluate the mean and standard deviation of emotion trajectories in event chains generated by PCR in comparison to mean emotion of events in isolation. Plots how the trajectories of the different emotions develop over the sentences of the narratives are shown in Figure 4. The colors used for the plots of the different emotions correspond to the ones used in Figure 3.

I. Correlation of Coherence and Emotion Uncertainty

Our analysis shows that the possibility of disambiguating events in emotion analysis depends on the particular emotion we want to trigger through contextual narratives. Now, we want to find out if emotion properties of the initial events already indicate whether disambiguation is possible through coherent narratives. Therefore, we investigate the relationship between model uncertainty in emotion analysis and the coherence of the generated backstories. We quantify uncertainty using the entropy of the predicted probability distribution.

To examine a potential correlation with the average coherence score, we calculate the Pearson ($r^{(p)}$) and Spearman ($r^{(s)}$) correlation of emotion model uncertainty e, i.e. the entropy of the probability distribution of zero-shot emotion analysis on the events, with average coherence c of the corresponding event chains. The correlation scores of events to the event chains generated through the various methods are as follows.

- Baseline: $r^{(p)} = -.09$ (**), $r^{(s)} = -.10$ (**)
- PC: $r^{(p)} = -.08$ (*), $r^{(s)} = -.11$ (***)
- PCR: $r^{(p)} = -.13$ (***), $r^{(s)} = -.16$ (***)

$e(c)$ / $e(b \cdot s_5)$	Anger	Boredom	Disgust	Fear	Guilt	Joy	Pride	Relief	Sadness	Shame	Surprise	Trust	No-Emot.	Σ
Anger	278	11	28	44	12	95	34	201	118	33	124	11	11	1000
Boredom	68	76	18	8	9	172	48	306	97	20	126	12	40	1000
Disgust	221	16	167	42	26	60	21	173	78	39	116	14	27	1000
Fear	55	2	5	288	18	81	28	264	64	42	120	18	15	1000
Guilt	63	1	13	38	253	88	50	232	70	102	78	5	7	1000
Joy	30	11	0	20	6	247	48	454	52	7	94	11	20	1000
Pride	19	4	3	16	6	239	240	308	33	24	81	13	14	1000
Relief	42	15	10	43	16	97	28	576	46	39	82	3	3	1000
Sadness	72	10	13	36	19	126	49	173	320	24	115	19	24	1000
Shame	79	3	9	34	110	68	46	189	105	230	105	6	16	1000
Surprise	54	8	6	19	10	95	18	312	45	22	387	6	18	1000
Trust	23	6	8	32	12	185	42	482	30	8	99	55	18	1000
No-Emotion	48	39	18	24	8	123	37	401	91	22	123	16	50	1000
Σ	1052	202	298	644	505	1676	689	4071	1149	612	1650	189	263	13000

Table 20: Prompted emotion vs. top predicted emotion for event chains (generation method: Baseline).

$e(c)$ / $e(b \cdot s_5)$	Anger	Boredom	Disgust	Fear	Guilt	Joy	Pride	Relief	Sadness	Shame	Surprise	Trust	No-Emot.	Σ
Anger	182	20	20	31	6	116	33	198	147	22	177	19	29	1000
Boredom	44	66	24	12	9	195	51	313	81	16	119	21	49	1000
Disgust	148	23	176	34	32	105	24	157	65	24	142	26	44	1000
Fear	44	16	10	151	22	156	28	259	63	29	158	20	44	1000
Guilt	21	4	6	24	163	169	57	284	67	57	110	12	26	1000
Joy	18	7	3	11	6	269	71	429	36	16	84	24	26	1000
Pride	16	0	3	15	3	284	161	360	27	6	83	16	26	1000
Relief	43	10	5	29	14	156	29	552	25	37	77	3	20	1000
Sadness	34	22	11	14	10	210	62	174	230	15	149	16	53	1000
Shame	25	3	3	27	83	112	62	285	86	160	111	12	31	1000
Surprise	37	30	11	27	8	158	20	361	51	9	246	20	22	1000
Trust	24	11	10	21	11	216	34	424	26	12	114	65	32	1000
No-Emotion	21	28	7	14	8	218	51	415	52	8	97	19	62	1000
Σ	657	240	289	410	375	2364	683	4211	956	411	1667	273	464	13000

Table 21: Prompted emotion vs. top predicted emotion for event chains (generation method: PC).

$e(c)$ / $e(b \cdot s_5)$	Anger	Boredom	Disgust	Fear	Guilt	Joy	Pride	Relief	Sadness	Shame	Surprise	Trust	No-Emot.	Σ
Anger	257	12	30	35	6	47	30	172	194	28	154	8	27	1000
Boredom	63	121	31	15	5	127	41	251	153	15	105	10	63	1000
Disgust	219	4	257	55	37	56	22	101	74	31	89	18	37	1000
Fear	44	5	12	292	22	72	17	261	72	51	101	10	41	1000
Guilt	20	3	6	46	260	74	19	270	84	105	79	7	27	1000
Joy	18	3	1	15	4	281	90	422	27	19	75	25	20	1000
Pride	10	0	0	11	2	291	246	292	25	5	81	14	23	1000
Relief	34	4	8	40	11	109	18	630	36	43	59	0	8	1000
Sadness	30	5	6	20	6	137	56	118	433	22	113	11	43	1000
Shame	17	1	3	43	108	55	27	248	102	280	92	6	18	1000
Surprise	29	19	13	21	4	123	12	349	61	9	313	24	23	1000
Trust	22	0	4	22	7	206	51	420	29	9	91	113	26	1000
No-Emotion	21	42	6	15	4	128	70	405	77	13	104	31	84	1000
Σ	784	219	377	630	476	1706	699	3939	1367	630	1456	277	440	13000

Table 22: Prompted emotion vs. top predicted emotion for event chains (generation method: PCR).

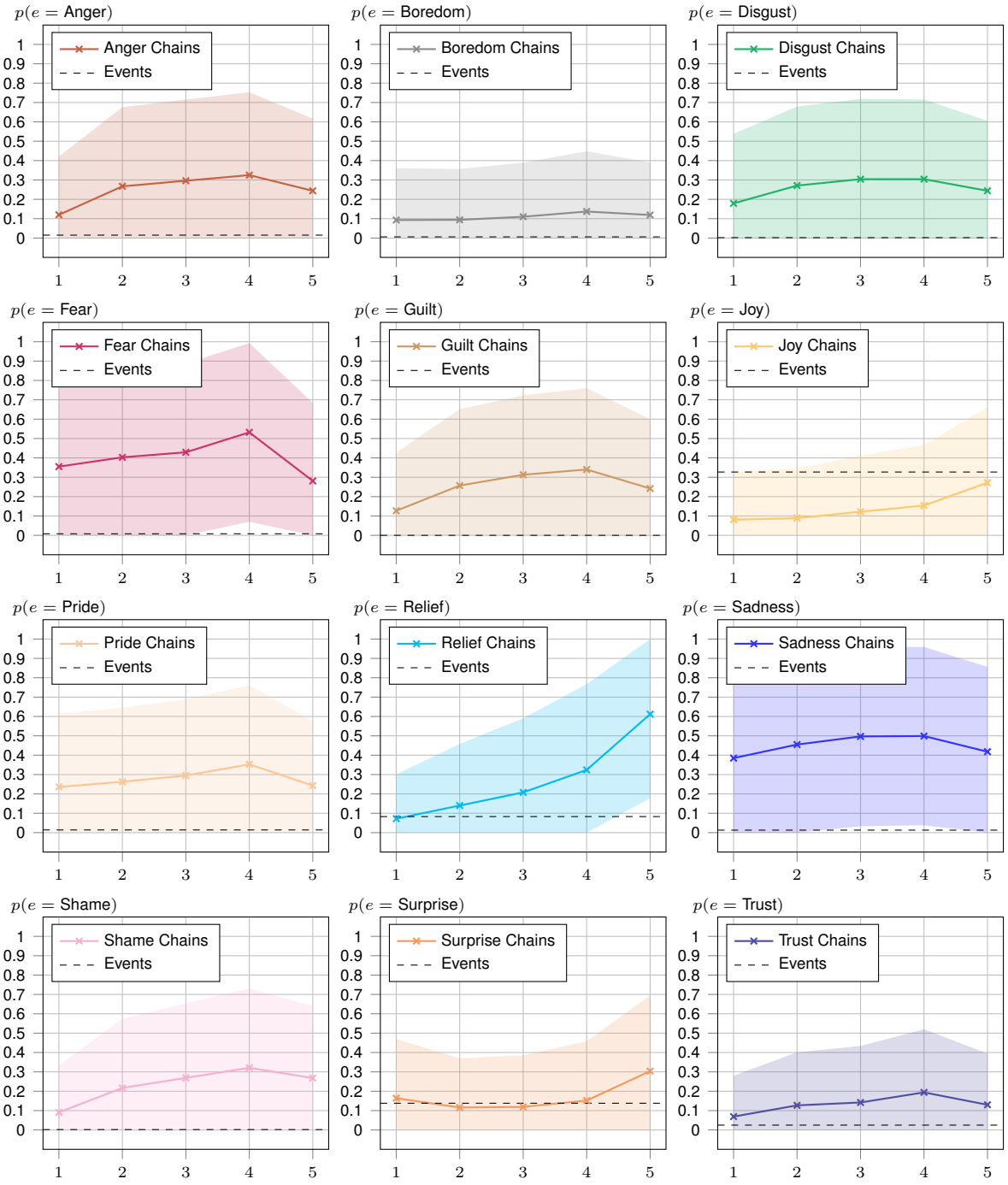

Figure 4: Mean and standard deviation of emotion trajectories in event chains (x-axes corresponding to the first n sentences of the chains; generation method: PCR) in comparison to mean emotion of events in isolation (dashed lines).

The results reveal a significant but modest negative correlation between the uncertainty of emotion predictions for the events and the coherence of the corresponding event chains. This finding implies that when the model exhibits uncertainty regarding the emotion impact of an event, such uncertainty may extend to the generation of narratives. Specifically, high entropy indicates that the event may lend itself to multiple interpretations or emotion responses. Consequently, narratives associated with such events may allow for variability in their placement within the storyline, leading to lower coherence scores assessed through the shuffle test. We therefore find that if the model shows low uncertainty in emotion analysis for events, this indicates that these events are easier to disambiguate using coherent narratives.